Proceedings of the 2009 Conference on Empirical
Methods in Natural Language Processing,
Singapore, August 2009

# A Unified Model of Phrasal and Sentential Evidence for Information Extraction

**Siddharth Patwardhan** and **Ellen Riloff**
School of Computing
University of Utah
Salt Lake City, UT 84112
{sidd,riloff}@cs.utah.edu

## Abstract

Information Extraction (IE) systems that extract role fillers for events typically look at the local context surrounding a phrase when deciding whether to extract it. Often, however, role fillers occur in clauses that are not directly linked to an event word. We present a new model for event extraction that jointly considers both the local context around a phrase along with the wider sentential context in a probabilistic framework. Our approach uses a *sentential event recognizer* and a *plausible role-filler recognizer* that is conditioned on event sentences. We evaluate our system on two IE data sets and show that our model performs well in comparison to existing IE systems that rely on local phrasal context.

## 1 Introduction

Information Extraction (IE) systems typically use extraction patterns (e.g., Soderland et al. (1995), Riloff (1996), Yangarber et al. (2000), Califf and Mooney (2003)) or classifiers (e.g., Freitag (1998), Freitag and McCallum (2000), Chieu et al. (2003), Bunescu and Mooney (2004)) to extract role fillers for events. Most IE systems consider only the immediate context surrounding a phrase when deciding whether to extract it. For tasks such as named entity recognition, immediate context is usually sufficient. But for more complex tasks, such as event extraction, a larger field of view is often needed to understand how facts tie together.

Most IE systems are designed to identify role fillers that appear as arguments to event verbs or nouns, either explicitly via syntactic relations or implicitly via proximity (e.g., *John murdered Tom* or *the murder of Tom by John*). But many facts are presented in clauses that do not contain event words, requiring discourse relations or deep structural analysis to associate the facts with event roles. For example, consider the sentences below:

***Seven people*** *have died*

*. . . and 30 were injured in India after terrorists launched an attack on the Taj Hotel.*

*. . . in Mexico City and its surrounding suburbs in a Swine Flu outbreak.*

*. . . after a tractor-trailer collided with a bus in Arkansas.*

***Two bridges*** *were destroyed*

*. . . in Baghdad last night in a resurgence of bomb attacks in the capital city.*

*. . . and $50 million in damage was caused by a hurricane that hit Miami on Friday.*

*. . . to make way for modern, safer bridges that will be constructed early next year.*

These examples illustrate a common phenomenon in text where information is not explicitly stated as filling an event role, but readers have no trouble making this inference. The role fillers above (***seven people***, ***two bridges***) occur as arguments to verbs that reveal state information (death, destruction) but are not event-specific (i.e., death and destruction can result from a wide variety of incident types). IE systems often fail to extract these role fillers because these systems do not recognize the immediate context as being relevant to the specific type of event that they are looking for.

We propose a new model for information extraction that incorporates both phrasal and sentential evidence in a unified framework. Our unified probabilistic model, called GLACIER, consists of two components: a model for *sentential event recognition* and a model for recognizing *plausible role fillers*. The Sentential Event Recognizer offers a probabilistic assessment of whether a sentence is discussing a domain-relevant event. The

Plausible Role-Filler Recognizer is then conditioned to identify phrases as role fillers based upon the assumption that the surrounding context is discussing a relevant event. This unified probabilistic model allows the two components to jointly make decisions based upon both the local evidence surrounding each phrase and the "peripheral vision" afforded by the sentential event recognizer.

This paper is organized as follows. Section 2 positions our research with respect to related work. Section 3 presents our unified probabilistic model for information extraction. Section 4 shows experimental results on two IE data sets, and Section 5 discusses directions for future work.

## 2 Related Work

Many event extraction systems rely heavily on the local context around words or phrases that are candidates for extraction. Some systems use extraction patterns (Soderland et al., 1995; Riloff, 1996; Yangarber et al., 2000; Califf and Mooney, 2003), which represent the immediate contexts surrounding candidate extractions. Similarly, classifier-based approaches (Freitag, 1998; Freitag and McCallum, 2000; Chieu et al., 2003; Bunescu and Mooney, 2004) rely on features in the immediate context of the candidate extractions. Our work seeks to incorporate additional context into IE.

Indeed, several recent approaches have shown the need for global information to improve IE performance. Maslennikov and Chua (2007) use discourse trees and local syntactic dependencies in a pattern-based framework to incorporate wider context. Finkel et al. (2005) and Ji and Grishman (2008) incorporate global information by enforcing event role or label consistency over a document or across related documents. In contrast, our approach simply creates a richer IE model for individual extractions by expanding the "field of view" to include the surrounding sentence.

The two components of the unified model presented in this paper are somewhat similar to our previous work (Patwardhan and Riloff, 2007), where we employ a relevant region identification phase prior to pattern-based extraction. In that work we adopted a pipeline paradigm, where a classifier identifies relevant sentences and only those sentences are fed to the extraction module. Our unified probabilistic model described in this paper does not draw a hard line between relevant and irrelevant sentences, but gently balances the influence of both local and sentential contexts through probability estimates.

## 3 A Unified IE Model that Combines Phrasal and Sentential Evidence

We introduce a probabilistic model for event-based IE that balances the influence of two kinds of contextual information. Our goal is to create a model that has the flexibility to make extraction decisions based upon strong evidence from the local context, or strong evidence from the wider context coupled with a more general local context. For example, some phrases explicitly refer to an event, so they almost certainly warrant extraction regardless of the wider context (e.g., *terrorists launched an attack*).[1] In contrast, some phrases are potentially relevant but too general to warrant extraction on their own (e.g., *people died* could be the result of different incident types). If we are confident that the sentence discusses an event of interest, however, then such phrases could be reliably extracted.

Our unified model for IE (GLACIER) combines two types of contextual information by incorporating it into a probabilistic framework. To determine whether a noun phrase instance $NP_i$ should be extracted as a filler for an event role, GLACIER computes the joint probability that $NP_i$:

(1) appears in an event sentence, and
(2) is a legitimate filler for the event role.

Thus, GLACIER is designed for noun phrase extraction and, mathematically, its decisions are based on the following joint probability:

$$P(EvSent(S_{NP_i}), PlausFillr(NP_i))$$

where $S_{NP_i}$ is the sentence containing noun phrase $NP_i$. This probability estimate is based on contextual features $F$ appearing within $S_{NP_i}$ and in the local context of $NP_i$. Including $F$ in the joint probability, and applying the product rule, we can split our probability into two components:

$$\begin{aligned} P(EvSent(S_{NP_i}), PlausFillr(NP_i)|F) = \\ P(EvSent(S_{NP_i})|F) \\ * \quad P(PlausFillr(NP_i)|EvSent(S_{NP_i}), F) \end{aligned}$$

These two probability components, in the expression above, form the basis of the two modules in

[1] There are always exceptions of course, such as hypothetical statements, but they are relatively uncommon.

our IE system – the *sentential event recognizer* and the *plausible role-filler recognizer*. In arriving at a decision to extract a noun phrase, our unified model for IE uses these modules to estimate the two probabilities based on the set of contextual features $F$. Note that having these two probability components allows the system to gently balance the influence from the sentential and phrasal contexts, without having to make hard decisions about sentence relevance or phrases in isolation.

In this system, the sentential event recognizer is embodied in the probability component $P(EvSent(S_{NP_i})|F)$. This is essentially the probability of a sentence describing a relevant event. Similarly, the plausible role-filler recognizer is embodied by the probability $P(PlausFillr(NP_i)|EvSent(S_{NP_i}), F)$. This component, therefore, estimates the probability that a noun phrase fills a specific event role, *assuming that the noun phrase occurs in an event sentence*. Many different techniques could be used to produce these probability estimates. In the rest of this section, we present the specific models that we used for each of these components.

### 3.1 Plausible Role-Filler Recognizer

The plausible role-filler recognizer is similar to most traditional IE systems, where the goal is to determine whether a noun phrase can be a legitimate filler for a specific type of event role based on its local context. Pattern-based approaches match the context surrounding a phrase using lexico-syntactic patterns or rules. However, most of these approaches do not produce probability estimates for the extractions. Classifier-based approaches use machine learning classifiers to make extraction decisions, based on features associated with the local context. Any classifier that can generate probability estimates, or similar confidence values, could be plugged into our model.

In our work, we use a Naïve Bayes classifier as our plausible role-filler recognizer. The probabilities are computed using a generative Naïve Bayes framework, based on local contextual features surrounding a noun phrase. These clues include lexical matches, semantic features, and syntactic relations, and will be described in more detail in Section 3.3. The Naïve Bayes (NB) plausible role-filler recognizer is defined as follows:

$$P(PlausFillr(NP_i)|EvSent(S_{NP_i}), F) = \frac{1}{Z} P(PlausFillr(NP_i)|EvSent(S_{NP_i})) * \prod_{f_i \in F} P(f_i|PlausFillr(NP_i), EvSent(S_{NP_i}))$$

where $F$ is the set of local contextual features and $Z$ is the normalizing constant. The prior $P(PlausFillr(NP_i)|EvSent(S_{NP_i}))$ is estimated from the fraction of role fillers in the training data. The product term in the equation is the likelihood, which makes the simplifying assumption that all of the features in $F$ are independent of one another. It is important to note that these probabilities are conditioned on the noun phrase $NP_i$ appearing in an event sentence.

Most IE systems need to extract several different types of role fillers for each event. For instance, to extract information about terrorist incidents a system may extract the names of perpetrators, victims, targets, and weapons. We create a separate IE model for each type of event role. To construct a unified IE model for an event role, we must specifically create a plausible role-filler recognizer for that event role, but we can use a single sentential event recognizer for all of the role filler types.

### 3.2 Sentential Event Recognizer

The task at hand for the sentential event recognizer is to analyze features in a sentence and estimate the probability that the sentence is discussing a relevant event. This is very similar to the task performed by text classification systems, with some minor differences. Firstly, we are dealing with the classification of sentences, as opposed to entire documents. Secondly, we need to generate a probability estimate of the "class", and not just a class label. Like the plausible role-filler recognizer, here too we employ machine learning classifiers to estimate the desired probabilities.

#### 3.2.1 Naïve Bayes Event Recognizer

Since Naïve Bayes classifiers estimate class probabilities, we employ such a classifier to create a sentential event recognizer:

$$P(EvSent(S_{NP_i})|F) = \frac{1}{Z} P(EvSent(S_{NP_i})) * \prod_{f_i \in F} P(f_i|EvSent(S_{NP_i}))$$

where $Z$ is the normalizing constant and $F$ is the set of contextual features in the sentence. The

prior $P(EvSent_{S(NP_i)})$ is obtained from the ratio of event and non-event sentences in the training data. The product term in the equation is the likelihood, which makes the simplifying assumption that the features in $F$ are independent of one another. The features used by the model will be described in Section 3.3.

A known issue with Naïve Bayes classifiers is that, even though their classification accuracy is often quite reasonable, their probability estimates are often poor (Domingos and Pazzani, 1996; Zadrozny and Elkan, 2001; Manning et al., 2008). The problem is that these classifiers tend to overestimate the probability of the predicted class, resulting in a situation where most probability estimates from the classifier tend to be either extremely close to 0.0 or extremely close to 1.0. We observed this problem in our classifier too, so we decided to explore an additional model to estimate probabilities for the sentential event recognizer. This second model, based on SVMs, is described next.

#### 3.2.2 SVM Event Recognizer

Given the all-or-nothing nature of the probability estimates that we observed from the Naïve Bayes model, we decided to try using a Support Vector Machine (SVM) (Vapnik, 1995; Joachims, 1998) classifier as an alternative to Naïve Bayes. One of the issues with doing this is that SVMs are not probabilistic classifiers. SVMs make classification decisions using on a *decision boundary* defined by *support vectors* identified during training. A decision function is applied to unseen test examples to determine which side of the decision boundary those examples lie. While the values obtained from the decision function only indicate class assignments for the examples, we used these values to produce confidence scores for our sentential event recognizer.

To produce a confidence score from the SVM classifier, we take the values generated by the decision function for each test instance and normalize them based on the minimum and maximum values produced across all of the test instances. This normalization process produces values between 0 and 1 that we use as a rough indicator of the confidence in the SVM's classification. We observed that we could effect a consistent recall/precision trade-off by using these values as thresholds for classification decisions, which suggests that this approach worked reasonably well for our task.

### 3.3 Contextual Features

We used a variety of contextual features in both components of our system. The plausible role-filler recognizer uses the following types of features for each candidate noun phrase $NP_i$: *lexical head* of $NP_i$, *semantic class* of $NP_i$'s lexical head, *named entity tags* associated with $NP_i$ and *lexico-syntactic patterns* that represent the local context surrounding $NP_i$. The feature set is automatically generated from the texts. Each feature is assigned a binary value for each instance, indicating either the presence or absence of the feature.

The *named-entity* features are generated by the freely available Stanford NER tagger (Finkel et al., 2005). We use the pre-trained NER model that comes with the software to identify person, organization and location names. The syntactic and semantic features are generated by the Sundance/AutoSlog system (Riloff and Phillips, 2004). We use the Sundance shallow parser to identify lexical heads, and use its semantic dictionaries to assign semantic features to words. The AutoSlog pattern generator (Riloff, 1996) is used to create the *lexico-syntactic pattern* features that capture local context around each noun phrase.

Our training sets produce a very large number of features, which initially bogged down our classifiers. Consequently, we reduced the size of the feature set by discarding all features that appeared four times or less in the training set.

Our sentential event recognizer uses the same contextual features as the plausible role-filler recognizer, except that features are generated for every NP in the sentence. In addition, it uses three types of sentence-level features: *sentence length*, *bag of words*, and *verb tense*, which are also binary features. We have two binary *sentence length* features indicating that the sentence is long (greater than 35 words) or is short (shorter than 5 words). Additionally, all of the words in each sentence in the training data are generated as *bag of words* features for the sentential model. Finally, we generate *verb tense* features from all verbs appearing in each sentence. Here too we apply a frequency cutoff and eliminate all features that appear four times or less in the training data.

## 4 IE Evaluation

### 4.1 Data Sets

We evaluated the performance of our IE system on two data sets: the MUC-4 terrorism corpus (Sund-

heim, 1992), and a ProMed disease outbreaks corpus (Phillips and Riloff, 2007; Patwardhan and Riloff, 2007). The MUC-4 data set is a standard IE benchmark collection of news stories about terrorist events. It contains 1700 documents divided into 1300 development (DEV) texts, and four test sets of 100 texts each (TST1, TST2, TST3, and TST4). Unless otherwise stated, our experiments adopted the same training/test split used in previous research: the 1300 DEV texts for training, 200 texts (TST1+TST2) for tuning, and 200 texts (TST3+TST4) as the blind test set. We evaluated our system on five MUC-4 string roles: *perpetrator individuals*, *perpetrator organizations*, *physical targets*, *victims*, and *weapons*.

The ProMed corpus consists of 120 documents obtained from ProMed-mail[2], a freely accessible global electronic reporting system for outbreaks of diseases. These 120 documents are paired with corresponding answer key templates. Unless otherwise noted, all of our experiments on this data set used 5-fold cross validation. We extracted two types of event roles: *diseases* and *victims*[3].

Unlike some other IE data sets, many of the texts in these collections do not describe a relevant event. Only about half of the MUC-4 articles describe a specific terrorist incident[4], and only about 80% of the ProMed articles describe a disease outbreak. The answer keys for the irrelevant documents are therefore empty. IE systems are especially susceptible to false hits when they can be given texts that contain no relevant events.

The complete IE task involves the creation of answer key templates, one template per incident (many documents in our data sets describe multiple events). Our work focuses on accurately extracting the facts from the text and not on template generation per se (e.g., we are not concerned with coreference resolution or which extraction belongs in which template). Consequently, our experiments evaluate the accuracy of the extractions individually. We used *head noun* scoring, where an extraction is considered to be correct if its head noun matches the head noun in the answer key.[5]

[2]http://www.promedmail.org

[3]The "victims" can be people, animals, or plants.

[4]With respect to the definition of terrorist incidents in the MUC-4 guidelines (Sundheim, 1992).

[5]Pronouns were discarded from both the system responses and the answer keys since we do not perform coreference resolution. Duplicate extractions (e.g., the same string extracted multiple times from the same document) were conflated before being scored, so they count as just one hit or one miss.

### 4.2 Baselines

We generated three baselines to use as comparisons with our IE system. As our first baseline, we used AutoSlog-TS (Riloff, 1996), which is a weakly-supervised, pattern-based IE system available as part of the Sundance/AutoSlog software package (Riloff and Phillips, 2004). Our previous work in event-based IE (Patwardhan and Riloff, 2007) also used a pattern-based approach that applied *semantic affinity* patterns to relevant regions in text. We use this system as our second baseline. As a third baseline, we trained a Naïve Bayes IE classifier that is analogous to the plausible role-filler recognizer in our unified IE model, except that this baseline system is not conditioned on the assumption of having an event sentence. Consequently, this baseline NB classifier is akin to a traditional supervised learning-based IE system that uses only local contextual features to make extraction decisions. Formally, the baseline NB classifier uses the formula:

$$P(PlausFillr(NP_i)|F) = \frac{1}{Z}P(PlausFillr(NP_i)) * \prod_{f_i \in F} P(f_i|PlausFillr(NP_i))$$

where $F$ is the set of local features, $P(PlausFillr(NP_i))$ is the prior probability, and $Z$ is the normalizing constant. We used the Weka (Witten and Frank, 2005) implementation of Naïve Bayes for this baseline NB system.

**New Jersey, February, 26.** An outbreak of Ebola has been confirmed in Mercer County, New Jersey. Five teenage boys appear to have contracted the deadly virus from an unknown source. The CDC is investigating the cases and is taking measures to prevent the spread...

| | |
|---|---|
| **Disease:** | Ebola |
| **Victims:** | Five teenage boys |
| **Location:** | Mercer County, New Jersey |
| **Date:** | February 26 |

Figure 1: A *Disease Outbreak* Event Template

Both the MUC-4 and ProMed data sets have separate answer keys rather than annotated source documents. Figure 1 shows an example of a document and its corresponding answer key template. To train the baseline NB system, we identify all instances of each answer key string in the source document and consider every instance a positive training example. This produces noisy training data, however, because some instances occur in

| | PerpInd | | | PerpOrg | | | Target | | | Victim | | | Weapon | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F | P | R | F | P | R | F |
| AutoSlog-TS | .33 | .49 | .40 | .52 | .33 | .41 | .54 | .59 | .56 | .49 | .54 | .51 | .38 | .44 | .41 |
| Sem Affinity | .48 | .39 | .43 | .36 | .58 | .45 | .56 | .46 | .50 | .46 | .44 | .45 | .53 | .46 | .50 |
| NB .50 | .36 | .34 | .35 | .35 | .46 | .40 | .53 | .49 | .51 | .50 | .50 | .50 | 1.00 | .05 | .10 |
| NB .70 | .41 | .25 | .31 | .43 | .31 | .36 | .58 | .42 | .48 | .58 | .37 | .45 | 1.00 | .04 | .07 |
| NB .90 | .51 | .17 | .25 | .56 | .15 | .24 | .67 | .30 | .41 | .75 | .23 | .36 | 1.00 | .02 | .04 |

Table 1: Baseline Results on MUC-4

| | Disease | | | Victim | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| AutoSlog-TS | .33 | .60 | .43 | .36 | .49 | .41 |
| Sem Affinity | .31 | .49 | .38 | .41 | .47 | .44 |
| NB .50 | .20 | .73 | .31 | .29 | .56 | .39 |
| NB .70 | .23 | .67 | .34 | .37 | .52 | .44 |
| NB .90 | .34 | .59 | .43 | .47 | .39 | .43 |

Table 2: Baseline Results on ProMed

undesirable contexts. For example, if the string "man" appears in an answer key as a victim, one instance of "man" may refer to the actual victim in an event sentence, while another instance of "man" may occur in a non-event context (e.g., background information) or may refer to a completely different person.

We report three evaluation metrics in our experiments: precision (P), recall (R), and F-score (F), where recall and precision are equally weighted. For the Naïve Bayes classifier, the natural threshold for distinguishing between positive and negative classes is 0.5, but we also evaluated this classifier with thresholds of 0.7 and 0.9 to see if we could effect a recall/precision trade-off. Tables 1 and 2 present the results of our three baseline systems. The NB classifier performs comparably to AutoSlog-TS and Semantic Affinity on most event roles, although a threshold of 0.90 is needed to reach comparable performance on ProMed. The relatively low numbers across the board indicate that these corpora are challenging, but these results suggest that our plausible role-filler recognizer is competitive with other existing IE systems. In Section 4.4 we will show how our unified IE model compares to these baselines. But before that (in the next section) we evaluate the quality of the second component of our IE system: the sentential event recognizer.

### 4.3 Sentential Event Recognizer Models

The sentential event recognizer is one of the core contributions of this research, so in this section we evaluate it by itself, before we employ it within the unified framework. The purpose of the sentential event recognizer is to determine whether a sentence is discussing a domain-relevant event. For our data sets, the classifier must decide whether a sentence is discussing a terrorist incident (MUC-4) or a disease outbreak (ProMed). Ideally, we want such a classifier to operate independently from the answer keys and the extraction task per se. For example, a terrorism IE system could be designed to extract only perpetrators and victims of terrorist events, or it could be designed to extract only targets and locations. The job of the sentential event recognizer remains the same: to identify sentences that discuss a terrorist event. How to train and evaluate such a system is a difficult question. In this section, we present two approaches that we explored to generate the training data: (a) using the IE answer keys, and (b) using human judgements.

#### 4.3.1 Sentence Annotation via Answer Keys

We have argued that the *event relevance* of a sentence should not be tied to a specific set of event roles. However, the IE answer keys can be used to identify some sentences that describe an event, because they contain an answer string. So we can map the answer strings back to sentences in the source documents to automatically generate event sentence annotations.[6] These annotations will be noisy, though, because an answer string can appear in a non-event sentence, and some event sentences may not contain any answer strings. The alternative, however, is sentence annotations by humans, which (as we will discuss in Section 4.3.2) is challenging.

#### 4.3.2 Sentence Annotation via Human Judgements

For many sentences there is a clear consensus among people that an event is being discussed. For example, most readers would agree that sentence (1) below is describing a terrorist event, while sen-

[6] A similar strategy was used in previous work (Patwardhan and Riloff, 2007) to generate a test set for the evaluation of a relevant region classifier.

| | | *Evaluation on Answer Keys* | | | | | | | *Evaluation on Human Annotations* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **Event** | | | **Non-Event** | | | | **Event** | | | **Non-Event** | | |
| | | Acc | Pr | Rec | F | Pr | Rec | F | Acc | Pr | Rec | F | Pr | Rec | F |
| ***MUC-4 (Terrorism)*** | | | | | | | | | | | | | | | |
| *Ans* | NB | .80 | .57 | .55 | .56 | .86 | .87 | .87 | .81 | .46 | .60 | .52 | .91 | .85 | .88 |
| | SVM | .80 | .68 | .42 | .52 | .84 | .93 | .88 | .83 | .55 | .44 | .49 | .88 | .91 | .90 |
| *Hum* | NB | .82 | .64 | .48 | .55 | .85 | .92 | .88 | **.85** | .56 | .57 | .57 | .91 | .91 | .91 |
| | SVM | .79 | .64 | .41 | .50 | .83 | .91 | .87 | .84 | .62 | .51 | .56 | .90 | .91 | .91 |
| ***ProMed (Disease Outbreaks)*** | | | | | | | | | | | | | | | |
| *Ans* | NB | .75 | .62 | .61 | .61 | .81 | .82 | .82 | .72 | .43 | .58 | .50 | .86 | .77 | .81 |
| | SVM | .74 | .78 | .31 | .44 | .74 | .95 | .83 | .76 | .51 | .26 | .35 | .80 | .92 | .86 |
| *Hum* | NB | .73 | .61 | .46 | .52 | .77 | .86 | .81 | **.79** | .56 | .57 | .56 | .87 | .86 | .86 |
| | SVM | .70 | .62 | .32 | .42 | .73 | .89 | .81 | **.79** | .62 | .42 | .50 | .84 | .90 | .87 |

Table 3: Sentential Event Recognizers Results (5-fold Cross-Validation)

| | *Evaluation on Human Annotations* | | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Event** | | | **Non-Event** | | |
| | Acc | Pr | Rec | F | Pr | Rec | F |
| NB | .83 | .50 | .70 | .58 | .94 | .86 | .90 |
| SVM | **.89** | .83 | .39 | .53 | .89 | .98 | .94 |

Table 4: Sentential Event Recognizer Results for MUC-4 using 1300 Documents for Training

tence (2) is not. However it is difficult to draw a clear line. Sentence (3), for example, describes an action taken in response to a terrorist event. Is this a terrorist event sentence? Precisely how to define an *event sentence* is not obvious.

> (1) *Al Qaeda operatives launched an attack on the Madrid subway system.*
>
> (2) *Madrid has a population of about 3.2 million people.*
>
> (3) *City officials stepped up security in response to the attacks.*

We tackled this issue by creating detailed annotation guidelines to define the notion of an event sentence, and conducting a human annotation study. The guidelines delineated a general time frame for the beginning and end of an event, and constrained the task to focus on specific incidents that were reported in the IE answer key. We gave the annotators a brief description (e.g., *murder in Peru*) of each event that had a filled answer key in the data set. They only labeled sentences that discussed those particular events.

We employed two human judges, who annotated 120 documents from the ProMed test set, and 100 documents from the MUC-4 test set. We asked both judges to label 30 of the same documents from each data set so that we could compute inter-annotator agreement. The annotators had an agreement of 0.72 Cohen's $\kappa$ on the ProMed data, and 0.77 Cohen's $\kappa$ on the MUC-4 data. Given the difficulty of this task, we were satisfied that this task is reasonably well-defined and the annotations are of good quality.

#### 4.3.3 Event Recognizer Results

We evaluated the two sentential event recognizer models described in Section 3.2 in two ways: (1) using the answer key sentence annotations for training/testing, and (2) using the human annotations for training/testing. Table 3 shows the results for all combinations of training/testing data. Since we only have human annotations for 100 MUC-4 texts and 120 ProMed texts, we performed 5-fold cross-validation on these documents. For our classifiers, we used the Weka (Witten and Frank, 2005) implementation of Naïve Bayes and the SVMLight (Joachims, 1998) implementation of the SVM. For each classifier we report overall accuracy, and precision, recall and F-scores with respect to both the positive and negative classes (event vs. non-event sentences).

The rows labeled *Ans* show the results for models trained via answer keys, and the rows labeled *Hum* show the results for the models trained with human annotations. The left side of the table shows the results using the answer key annotations for evaluation, and the right side of the table shows the results using the human annotations for evaluation. One expects classifiers to perform best when they are trained and tested on the same type of data, and our results bear this out – the classifiers that were trained and tested on the same kind of annotations do best. The boldfaced numbers represent the best accuracies achieved for each domain. As we would expect, the classifiers that are both trained and tested with human annotations (*Hum*) show the best performance, with the Naïve Bayes achieving the best accuracy of 85% on the

| | PerpInd | | | PerpOrg | | | Target | | | Victim | | | Weapon | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F | P | R | F | P | R | F |
| AutoSlog-TS | .33 | .49 | .40 | .52 | .33 | .41 | .54 | .59 | **.56** | .49 | .54 | .51 | .38 | .44 | .41 |
| Sem Affinity | .48 | .39 | .43 | .36 | .58 | **.45** | .56 | .46 | .50 | .46 | .44 | .45 | .53 | .46 | .50 |
| NB (baseline) | .36 | .34 | .35 | .35 | .46 | .40 | .53 | .49 | .51 | .50 | .50 | .50 | 1.00 | .05 | .10 |
| GLACIER | | | | | | | | | | | | | | | |
| NB/NB .90 | .39 | .59 | .47 | .33 | .51 | .40 | .39 | .72 | .51 | .52 | .54 | .53 | .47 | .55 | .51 |
| NB/SVM .40 | .51 | .58 | .54 | .34 | .45 | .38 | .42 | .72 | .53 | .55 | .58 | **.56** | .57 | .53 | **.55** |
| NB/SVM .50 | .66 | .47 | **.55** | .41 | .26 | .32 | .50 | .62 | .55 | .62 | .36 | .45 | .64 | .43 | .52 |

Table 5: Unified IE Model on MUC-4

MUC-4 texts, and the SVM achieving the best accuracy of 79% on the ProMed texts.

The recall and precision for non-event sentences is much higher than for event sentences. This classifier is forced to draw a hard line between the event and non-event sentences, which is a difficult task even for people. One of the advantages of our unified IE model, which will be described in the next section, is that it does not require hard decisions but instead uses a probabilistic estimate of how "event-ish" a sentence is.

Table 3 showed that models trained on human annotations outperform models trained on answer key annotations. But with the MUC-4 data, we have the luxury of 1300 training documents with answer keys, while we only have 100 documents with human annotations. Even though the answer key annotations are noisier, we have 13 times as much training data.

So we trained another sentential event recognizer using the entire MUC-4 training set. These results are shown in Table 4. Observe that using this larger (albeit noisy) training data does not appear to affect the Naïve Bayes model very much. Compared with the model trained on 100 manually annotated documents, its accuracy decreases by 2% from 85% to 83%. The SVM model, on the other hand, achieves an 89% accuracy when trained with the larger MUC-4 training data, compared to 84% accuracy for the model trained from the 100 manually labeled documents. Consequently, the sentential event recognizer models used in our unified IE framework (described in Section 4.4) are trained with this 1300 document training set.

### 4.4 Evaluation of the Unified IE Model

We now evaluate the performance of our unified IE model, GLACIER, which allows a plausible role-filler recognizer and a sentential event recognizer to make joint decisions about phrase extractions. Tables 5 and 6 present the results of the unified

| | Disease | | | Victim | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| AutoSlog-TS | .33 | .60 | .43 | .36 | .49 | .41 |
| Sem Affinity | .31 | .49 | .38 | .41 | .47 | **.44** |
| NB (baseline) | .34 | .59 | .43 | .47 | .39 | .43 |
| GLACIER | | | | | | |
| NB/NB .90 | .41 | .61 | **.49** | .38 | .52 | **.44** |
| NB/SVM .40 | .31 | .66 | .42 | .32 | .55 | .41 |
| NB/SVM .50 | .38 | .54 | .44 | .42 | .47 | **.44** |

Table 6: Unified IE Model on ProMed

IE model on the MUC-4 and ProMed data sets. The NB/NB systems use Naïve Bayes models for both components, while the NB/SVM systems use a Naïve Bayes model for the plausible role-filler recognizer and an SVM for the sentential event recognizer. As with our baseline system, we obtain good results using a threshold of 0.90 for our NB/NB model (i.e., only NPs with probability $\geq$ 0.90 are extracted). For our NB/SVM models, we evaluated using the default threshold (0.50) but observed that recall was sometimes low. So we also use a threshold of 0.40, which produces superior results. Here too, we used the Weka (Witten and Frank, 2005) implementation of the Naïve Bayes model and the SVMLight (Joachims, 1998) implementation of the SVM.

For the MUC-4 data, our unified IE model using the SVM (0.40) outperforms all 3 baselines on three roles (**PerpInd, Victim, Weapon**) and outperforms 2 of the 3 baselines on the **Target** role. When GLACIER outperforms the other systems it is often by a wide margin: the F-score for **PerpInd** jumped from 0.43 for the best baseline (Sem Affinity) to 0.54 for GLACIER, and the F-scores for **Victim** and **Weapon** each improved by 5% over the best baseline. These gains came from both increased recall and increased precision, demonstrating that GLACIER extracts some information that was missed by the other systems and is also less prone to false hits.

Only the **PerpOrg** role shows inferior performance. Organizations perpetrating a terrorist

event are often discussed later in a document, far removed from the main event description. For example, a statement that *Al Qaeda* is believed to be responsible for an attack would typically appear after the event description. As a result, the sentential event recognizer tends to generate low probabilities for such sentences. We believe that addressing this issue would require the use of discourse relations or the use of even larger context sizes. We intend to explore these avenues of research in future work.

On the ProMed data, GLACIER produces results that are similar to the baselines for the **Victim** role, but it outperforms the baselines for the **Disease** role. We find that for this domain, the unified IE model with the Naïve Bayes sentential event recognizer is superior to the unified IE model with the SVM classifier. For the **Disease** role, the F-score jumped 6%, from 0.43 for the best baseline systems (AutoSlog-TS and the NB baseline) to 0.49 for $\text{GLACIER}_{\text{NB/NB}}$. In contrast to the MUC-4 data, this improvement was mostly due to an increase in precision (up to 0.41), indicating that our unified IE model was effective at eliminating many false hits. For the **Victim** role, the performance of the unified model is comparable to the baselines. On this event role, the F-score of $\text{GLACIER}_{\text{NB/NB}}$ (0.44) matches that of the best baseline system (Sem Affinity, with 0.44). However, note that $\text{GLACIER}_{\text{NB/NB}}$ can achieve a 5% gain in recall over this baseline, at the cost of a 3% precision loss.

### 4.5 Specific Examples

Figure 2 presents some specific examples of extractions that are failed to be extracted by the baseline models, but are correctly identified by GLACIER because of its use of sentential evidence. Observe that in each of these examples, GLACIER correctly extracts the underlined phrases, in spite of the inconclusive evidence in the local contexts around them. In the last sentence in Figure 2, for example, GLACIER correctly makes the inference that the policemen in the bus (which was traveling on the bridge) are likely the victims of the terrorist event. Thus, we see that our system manages to balance the influence of the two probability components to make extraction decisions that would be impossible to make by relying only on the local phrasal context. In addition, the sentential event recognizer can also help improve precision by preventing extractions from non-event sentences.

THE MNR REPORTED ON 12 JANUARY THAT HEAVILY ARMED MEN IN CIVILIAN CLOTHES HAD INTERCEPTED A VEHICLE WITH OQUELI AND FLORES ENROUTE FOR LA AURORA AIRPORT AND THAT THE TWO POLITICAL LEADERS HAD BEEN KIDNAPPED AND WERE REPORTED MISSING.

**PerpInd**: HEAVILY ARMED MEN

---

THE SCANT POLICE INFORMATION SAID THAT THE DEVICES WERE APPARENTLY LEFT IN FRONT OF THE TWO BANK BRANCHES MINUTES BEFORE THE CURFEW BEGAN FOR THE 6TH CONSECUTIVE DAY – PRECISELY TO COUNTER THE WAVE OF TERRORISM CAUSED BY DRUG TRAFFICKERS.

**Weapon**: THE DEVICES

---

THOSE WOUNDED INCLUDE THREE EMPLOYEES OF THE GAS STATION WHERE THE CAR BOMB WENT OFF AND TWO PEOPLE WHO WERE WALKING BY THE GAS STATION AT THE MOMENT OF THE EXPLOSION.

**Victim**: THREE EMPLOYEES OF THE GAS STATION
**Victim**: TWO PEOPLE

---

MEMBERS OF THE BOMB SQUAD HAVE DEACTIVATED A POWERFUL BOMB PLANTED AT THE ANDRES AVELINO CACERES PARK, WHERE PRESIDENT ALAN GARCIA WAS DUE TO PARTICIPATE IN THE COMMEMORATION OF THE BATTLE OF TARAPACA.

**Victim**: PRESIDENT ALAN GARCIA

---

EPL [POPULAR LIBERATION ARMY] GUERRILLAS BLEW UP A BRIDGE AS A PUBLIC BUS, IN WHICH SEVERAL POLICEMEN WERE TRAVELING, WAS CROSSING IT.

**Victim**: SEVERAL POLICEMEN

Figure 2: Examples of GLACIER Extractions

## 5 Conclusions

We presented a unified model for IE that balances the influence of sentential context with local contextual evidence to improve the performance of event-based IE. Our experimental results showed that using sentential contexts indeed produced better results on two IE data sets. Our current model uses supervised learning, so one direction for future work is to adapt the model for weakly supervised learning. We also plan to incorporate discourse features and investigate even wider contexts to capture broader discourse effects.

## Acknowledgments

This work has been supported in part by the Department of Homeland Security Grant N0014-07-1-0152. We are grateful to Nathan Gilbert and Adam Teichert for their help with the annotation of event sentences.

## References


R. Bunescu and R. Mooney. 2004. Collective Information Extraction with Relational Markov Networks. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 438–445, Barcelona, Spain, July.

M. Califf and R. Mooney. 2003. Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction. *Journal of Machine Learning Research*, 4:177–210, December.

H. Chieu, H. Ng, and Y. Lee. 2003. Closing the Gap: Learning-Based Information Extraction Rivaling Knowledge-Engineering Methods. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 216–223, Sapporo, Japan, July.

P. Domingos and M. Pazzani. 1996. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 105–112, Bari, Italy, July.

J. Finkel, T. Grenager, and C. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, Ann Arbor, MI, June.

D. Freitag and A. McCallum. 2000. Information Extraction with HMM Structures Learned by Stochastic Optimization. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 584–589, Austin, TX, August.

D. Freitag. 1998. Toward General-Purpose Learning for Information Extraction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 404–408, Montreal, Quebec, August.

H. Ji and R. Grishman. 2008. Refining Event Extraction through Cross-Document Inference. In *Proceedings of ACL-08: HLT*, pages 254–262, Columbus, OH, June.

T. Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the Tenth European Conference on Machine Learning*, pages 137–142, April.

C. Manning, P. Raghavan, and H Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY.

M. Maslennikov and T. Chua. 2007. A Multi-resolution Framework for Information Extraction from Free Text. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 592–599, Prague, Czech Republic, June.

S. Patwardhan and E. Riloff. 2007. Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 717–727, Prague, Czech Republic, June.

W. Phillips and E. Riloff. 2007. Exploiting Role-Identifying Nouns and Expressions for Information Extraction. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pages 165–172, Borovets, Bulgaria, September.

E. Riloff and W. Phillips. 2004. An Introduction to the Sundance and AutoSlog Systems. Technical Report UUCS-04-015, School of Computing, University of Utah.

E. Riloff. 1996. Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Articial Intelligence*, pages 1044–1049, Portland, OR, August.

S. Soderland, D. Fisher, J. Aseltine, and W. Lehnert. 1995. CRYSTAL: Inducing a Conceptual Dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1314–1319, Montreal, Canada, August.

B. Sundheim. 1992. Overview of the Fourth Message Understanding Evaluation and Conference. In *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, pages 3–21, McLean, VA, June.

V. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer, New York, NY.

I. Witten and E. Frank. 2005. *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan–Kaufmann, San Francisco, CA.

R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen. 2000. Automatic Acquisition of Domain Knowledge for Information Extraction. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 940–946, Saarbrücken, Germany, August.

B. Zadrozny and C. Elkan. 2001. Obtaining Calibrated Probability Estimates from Decision Trees and Naïve Bayesian Classiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 609–616, Williamstown, MA, June.